Semantic and Knowledge Engineering Using ENVRI RM

Paul Martin[1], Xiaofeng Liao[1], Barbara Magagna[2], Markus Stocker[3,4], and Zhiming Zhao[1(✉)]

[1] Multiscale Networked Systems, University of Amsterdam, 1098XH Amsterdam, The Netherlands
pwmartin.research@gmail.com, {x.liao,z.zhao}@uva.nl
[2] Environment Agency Austria, Vienna, Austria
barbara.magagna@umweltbundesamt.at
[3] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
markus.stocker@tib.eu
[4] MARUM Center for Marine Environmental Sciences, PANGAEA Data Publisher for Earth & Environmental Science, Leobener Strasse 8, 28359 Bremen, Germany

Abstract. The ENVRI Reference Model provides architects and engineers with the means to describe the architecture and operational behaviour of environmental and Earth science research infrastructures (RIs) in a standardised way using the standard terminology. This terminology and the relationships between specific classes of concept can be used as the basis for the machine-actionable specification of RIs or RI subsystems.

Open Information Linking for Environmental RIs (OIL-E) is a framework for capturing architectural and design knowledge about environmental and Earth science RIs intended to help harmonise vocabulary, promote collaboration and identify common standards and technologies across different research infrastructure initiatives. At its heart is an ontology derived from the ENVRI Reference Model. Using this ontology, RI descriptions can be published as linked data, allowing discovery, querying and comparison using established Semantic Web technologies. It can also be used as an upper ontology by which to connect descriptions of RI entities (whether they be datasets, equipment, processes, etc.) that use other, more specific terminologies.

The ENVRI Knowledge Base uses OIL-E to capture information about environmental and Earth science RIs in the ENVRI community for query and comparison. The Knowledge Base can be used to identify the technologies and standards used for particular activities and services and as a basis for evaluating research infrastructure subsystems and behaviours against certain criteria, such as compliance with the FAIR data principles.

Keywords: Ontology · Knowledge base · Research infrastructure · Reference model

© The Author(s) 2020
Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 100–119, 2020.
https://doi.org/10.1007/978-3-030-52829-4_6

1 Introduction

The ENVRI Reference Model[1] (ENVRI RM) provides a standard set of stereotypes for the different classes of actor, information object, behaviour, etc. found within environmental and Earth science research infrastructures (RIs) [1, 2]. These stereotypes were derived from the study of the RIs participating in the ENVRI community cluster for environmental RIs in Europe[2]. ENVRI RM places all of these stereotypes in the context of the research data lifecycle, identifying the critical elements needed to facilitate data acquisition, curation, publishing, processing and use by a community of researchers in the environmental and Earth sciences, though many stereotypes are applicable more broadly to research infrastructure in general. By referring to the model, RI architects can identify the elements that are most important to them, determine any gaps within their own (planned) infrastructure, and compare against other RI specifications—in particular allowing them to look at how other RIs solved the same problems, and what technologies and standards they used to do so. Given the instantiation of ENVRI RM for a particular RI however, there is still the question of how the resulting information can be published in a way that is useful to as broad an audience as possible. For example, in addition to published documentation describing the modelling of a particular RI (with in-depth textual explanations and diagrams of major subsystems and their organisation and construction), it would also be convenient to be able to translate those documents into a form that can be programmatically queried and compared against other RI models in a systematic way.

This chapter describes how ENVRI RM was used as a basis to create Open Information Linking for Environmental Research Infrastructures (OIL-E) [3], a multi-view machine-readable ontology for describing RIs based on ENVRI RM that can act as an upper ontology for describing different entities and activities attributable to environmental and Earth science RIs. OIL-E was intended to:

1. Capture the terminology of ENVRI RM as a controlled resource for use in the annotation of RI documentation and other semantic enrichment activities.
2. Permit the translation of specific RI models produced using ENVRI RM into machine-readable RDF data that can be stored in a suitable knowledge base.
3. Assist in the association of other semantic descriptions for data, services and other RI elements with one another by acting as a ‘connective ontology’ for environmental and Earth science RI entity specifications.

As part of the second objective, in particular, an ENVRI Knowledge Base[3] has been under development to serve as an online information corpus about the ENVRI cluster of environmental science RIs. The ENVRI Knowledge Base gathers information collected about RI design and RI resources, structured according to the OIL-E ontology (and ENVRI RM) and provides access based on established Semantic Web technologies. It thus serves as a practical demonstrator of the kind of semantic search and query that OIL-E can facilitate.

[1] http://envri.eu/rm/.
[2] https://envri.eu/.
[3] http://kb.oil-e.net/.

In Sect. 2, we examine more closely the background and motivation behind using ENVRI RM to develop semantic and knowledge resources for the environmental and Earth science RI community. We describe the methodology applied in developing the OIL-E ontology (Sect. 3), and how we applied it to the modelling of RIs in the ENVRI cluster (Sect. 4). We then move on to discuss the ENVRI Knowledge Base (Sect. 5). Finally, we discuss where further development is needed or desired (Sect. 6) before drawing our conclusions (Sect. 7).

2 Background and Motivation

Environmental research increasingly depends on the collection and analysis of large volumes of data gathered from various sources including field observations, sensor networks, laboratory experiments and simulations based on expert models. Societal challenges facing the world today like climate change, food security and disaster prediction/response can only be addressed by making optimal use of such data, which also requires scientists to collaborate across disciplinary boundaries, as these challenges are intrinsically transdisciplinary in nature. Environmental and Earth science RIs support researchers in their interactions with a host of different data sources and analytical tools by providing access to combined corpora of curated research datasets via unified services and data portals, but no one RI fully encompasses the full research ecosystem [4], each typically serving a specific environmental domain or catering for a specific class of data. The challenge, therefore, is to functionally integrate existing environmental RIs to permit researchers to freely and effectively interact with the full range of research assets potentially available to them, allowing them to collaborate and conduct innovative interdisciplinary research regardless of the particular research community to which they belong. Realising this ideal requires a broad understanding of the fundamental commonalities of environmental science research infrastructure services, however: in terms of concepts, in terms of processes, in terms of data and services, and in terms of technology adoption. The process of achieving this understanding can be expedited by the use of a standard reference model (e.g. ENVRI RM), which can be used to construct formal descriptions of RIs and their major component elements.

ENVRI RM was constructed using the Reference Model for Open Distributed Processing (ODP) [5] for modelling complex distributed systems. ODP requires the modelling of a system from five different viewpoints (enterprise, information, computation, engineering and technology), with the correspondences between the five resulting views ensuring their mutual validity. This viewpoint-based approach provides clarity to each 'facet' of the end model by reducing the number of competing elements to only those that match a particular set of concerns (such as the flow of information through the system), while still retaining the aggregate complexity needed to model any substantive distributed system. ENVRI RM provides the five views prescribed by ODP (renaming the enterprise view as the *science* view in light of its subject) specialised for the common elements of environmental and Earth science RIs, as revealed in the study of participating RIs in Europe:

- The **science** viewpoint, which considers the main behaviours facilitated by a RI and the communities and resources involved in those behaviours.

- The **information** viewpoint, which identifies the information objects handled by a RI and their various states throughout the operation of the RI.
- The **computational** viewpoint, which identifies the logical computational elements that interact to support various RI operations.
- The **engineering** viewpoint, which describes how computational elements are distributed in an infrastructure, and the communication channels between infrastructural nodes.
- The **technology** viewpoint, which identifies the software, hardware and standards used to implement data and computational entities in a RI.

ENVRI RM uses three of the five views prescribed by ODP to capture the generic aspects common across all RIs (those being the science, information and computational views), and then uses the engineering and technology viewpoints to explore the more specific solutions and design patterns observed as being used by current RIs for the generic components prescribed in the three former views. Each view has its own concerns, and parts of those concerns may correspond to concerns in other views (for example information in one view may be used by computational elements in another); each view is thus able to describe particular key RI activities. For example, in Fig. 1, we show the components prescribed for raw data collection in the computational view as a UML component diagram. A *data transfer service* provides a *raw data collector* which brokers the streaming of data from an instrument (represented by an *instrument controller* computational object) to a data store (represented by a *data store controller*), with a *persistent identifier service* invoked to acquire an identifier for the resulting dataset and that dataset's existence registered with a *catalogue service*. For any given RI, these components are expected to be present in some form; perhaps the data collector is not a distinct component from the data transfer service, and perhaps the PID service is only invoked for certain types of data, but most actual cases of raw data collection should be describable in terms of this interaction template.

The use of methodologies such as ODP [6] helps guide the software engineering process by recognising the existence of different kinds of stakeholder in system development with different primary concerns and providing a multi-faceted modelling context that addresses each while maintaining an overall coherent specification. This benefits all parties by providing distinct specifications of each facet of the system that is sufficiently revealing the key characteristics of the system from one perspective. Simultaneously, these specifications can ignore details that are less relevant to that perspective, as long as those details are made evident in at least one of the other views so that they are not neglected by the combined specification. A similar benefit can be obtained in the design of ontologies and other formal models, where simple decompositions of systems with a particular perspective in mind often produce the most useful and easy to apply models. Conversely, trying to do 'too much' within the framework of a single ontology can make it more difficult to use and more likely to contain errors or controversies. We can instead create a linked set of interconnected ontologies (or partition a larger ontology into parallel-connected sub-ontologies with independent hierarchies); each sub-ontology then represents a different viewpoint, but with links to corresponding concepts in the other ontologies. This allows for more complex systems to be modelled while retaining the clarity of a simple yet serviceable ontology for each viewpoint. Such an approach

also provides the option to focus on a case-by-case basis on modelling those specific views deemed most useful, ignoring the other viewpoints not applicable to modellers' immediate concerns.

OIL-E is intended to provide such a multi-view framework for the modelling of environmental and Earth science RIs. The OIL-E ontology captures all stereotypes defined by ENVRI RM along with their essential relations and distributes them across the ENVRI RM viewpoints while also adding more cross-view relations to better facilitate classification and validation of RI models described using the ontology. For example, OIL-E allows for technologies (including both software and standards) used within a RI to be linked directly to the information objects and computational services that implement or use them. Using the stereotypes of ENVRI RM to produce a high-level, 'connective' ontology for RI specifications, OIL-E can provide a means for describing and maintaining constellations of loosely-coupled views on the same RI system, where the correspondences between concepts in different views might be difficult to express with complete precision, making the conception of a single canonical representation that integrates the full scope of all views difficult or intractable.

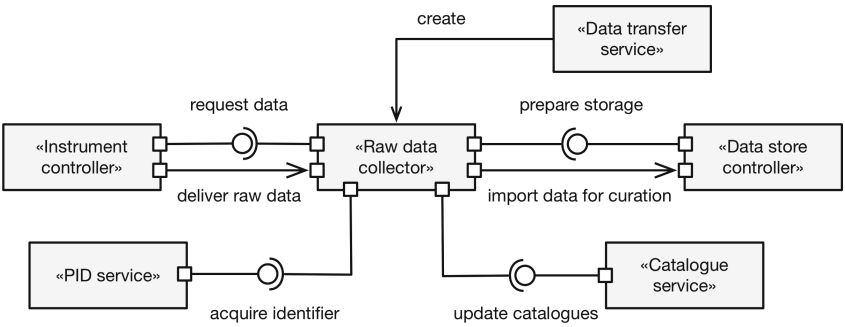


Fig. 1. A computational view of raw data acquisition: ENVRI RM specifies components and activities using UML (in this case, a component diagram).

3 Methodology

The ENVRI semantic linking framework was developed based on ENVRI RM. Open Information Linking for Environmental RIs (OIL-E) was designed to provide an upper ontology for RI descriptions based on ENVRI RM that can be used to contextualise different kinds of RI assets from an architectural or operational perspective. This is in contrast to being a general-purpose ontology for describing scientific phenomena like ENVO [7] or BFO [8]; OIL-E has more in common with conceptual models such as CERIF [9] that focus on the products and tools of research rather than on scientific classification itself, albeit more concerned with providing a controlled vocabulary for environmental science RIs in particular.

The multi-viewpoint approach intrinsic to ENVRI RM and inherited from ODP informs the design of OIL-E in many ways. Most notably, each viewpoint essentially

provides its own micro-ontology, with instances of the concepts defined that can then be related to concepts in other views via correspondences. Correspondences, as defined by ODP, describe relationships between entities existing in different views, and are used to anchor the different views with one another to ensure a coherent description of the same system. This allows OIL-E to operate as a 'hub' ontology, whereby specifications created in one view (e.g. information) can be used to dictate requirements on another view (e.g. computation). For example, given the specification of an information action to produce newly processed data from a persistent dataset, there must be an accompanying computational operation to carry out that action. Likewise, given a behaviour by which a researcher processes such data, there must be a computational service on which that operation can be invoked. It is also possible to extend each view using other, more specific ontologies (e.g. for describing datasets in the information view), which then inherit the relationships with concepts in the other views.

As a Semantic Web [10] ontology, OIL-E is written in OWL 2.0 [11] and published online[4], with the ontology itself split into two parts. The full ontology:

- Captures notions of research infrastructure from multiple perspectives: social infrastructure, physical research infrastructure (i.e. sites, observatories and devices) and computational infrastructure being the most evident.
- Clearly separates these different views on infrastructure, and then establishes their correspondences.
- Captures the most significant interactions between different actors and resources, and the information that is produced by such interactions.
- Helps establish the relationships between other existing standards and vocabularies in terms of the facets of infrastructure, infrastructure assets and infrastructure activity to which they apply.

The foundation of OIL-E is the **oil-base** base ontology, which provides a set of abstract concept classes derived from the most common elements observed in the ENVRI RM and distributed across the five standard ODP views. The purpose of **oil-base** is to capture the generic concepts not specific to environmental science RIs, and to act as a simple upper ontology for all further OIL-E extensions. Despite its application to research infrastructure, **oil-base** is not a general-purpose upper ontology for describing scientific phenomena, but rather is a means to gather architectural and procedural concepts used in a complex system, distribute them across the most appropriate views, and then model the correspondences across those views.

Figure 2 illustrates the core concept hierarchy and its subdivision into the top-level concepts for each ODP view. These concepts generally refer either to *objects* of discourse, *activities* involving such objects, or *attributes* of objects and activities. This simple categorisation is used as the basis for defining exclusivity and restrictions on object properties, as well as allowing certain concepts to exist in multiple views and many generic properties to be defined for use in multiple views (or across views). The separation of specific concepts to specific views is then done via inference using *classifier* concepts

[4] http://www.oil-e.net/.

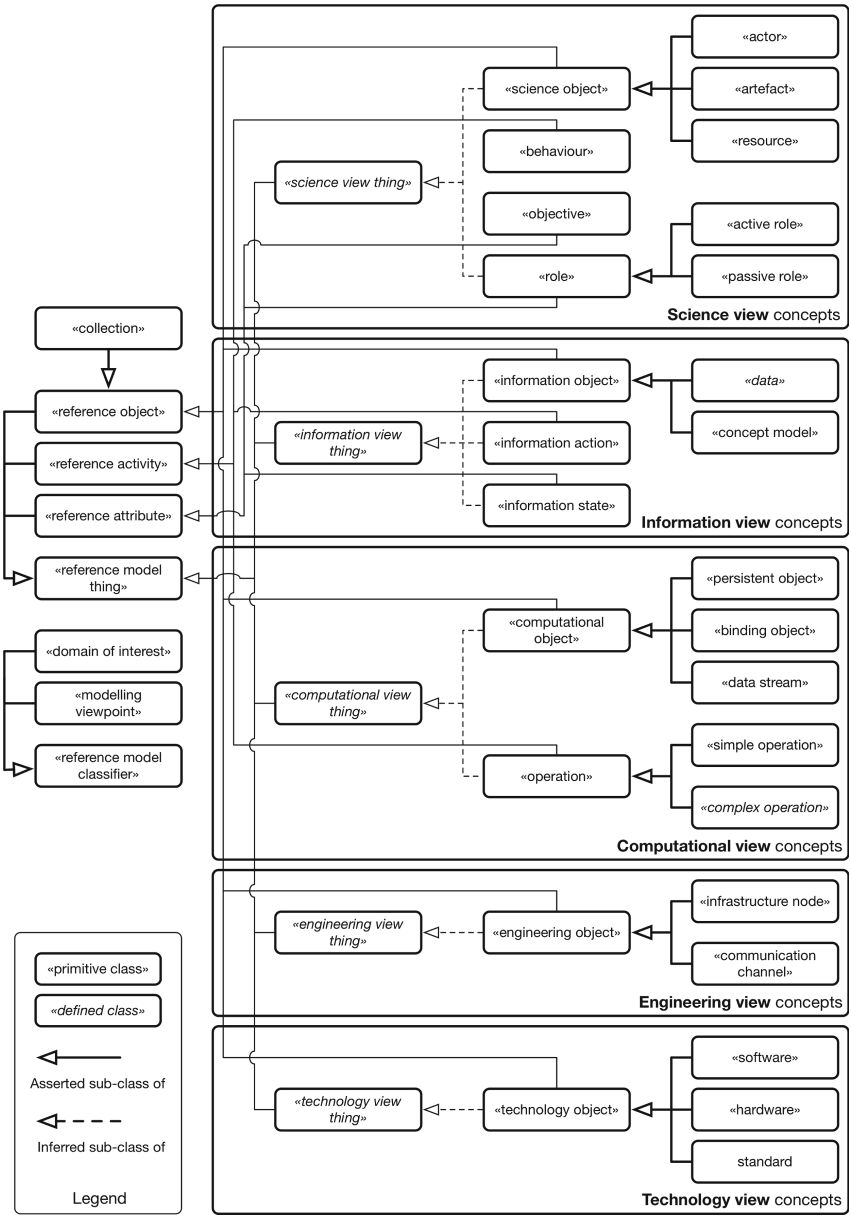


Fig. 2. The top-level concept hierarchy of each viewpoint in **oil-base**. Some sub-concepts have been omitted for brevity.

for which there are default definitions for each of the five ODP viewpoints. This has been done to make it easier to specify alternative viewpoints (e.g. a virtualisation viewpoint or a privacy viewpoint) should the original five ODP viewpoints be deemed insufficient to future modelling needs without requiring a substantial restructuring of the ontology. This

approach also minimises the number of concept classes that are derived from multiple parent classes, in line with standard ontology design best practices.

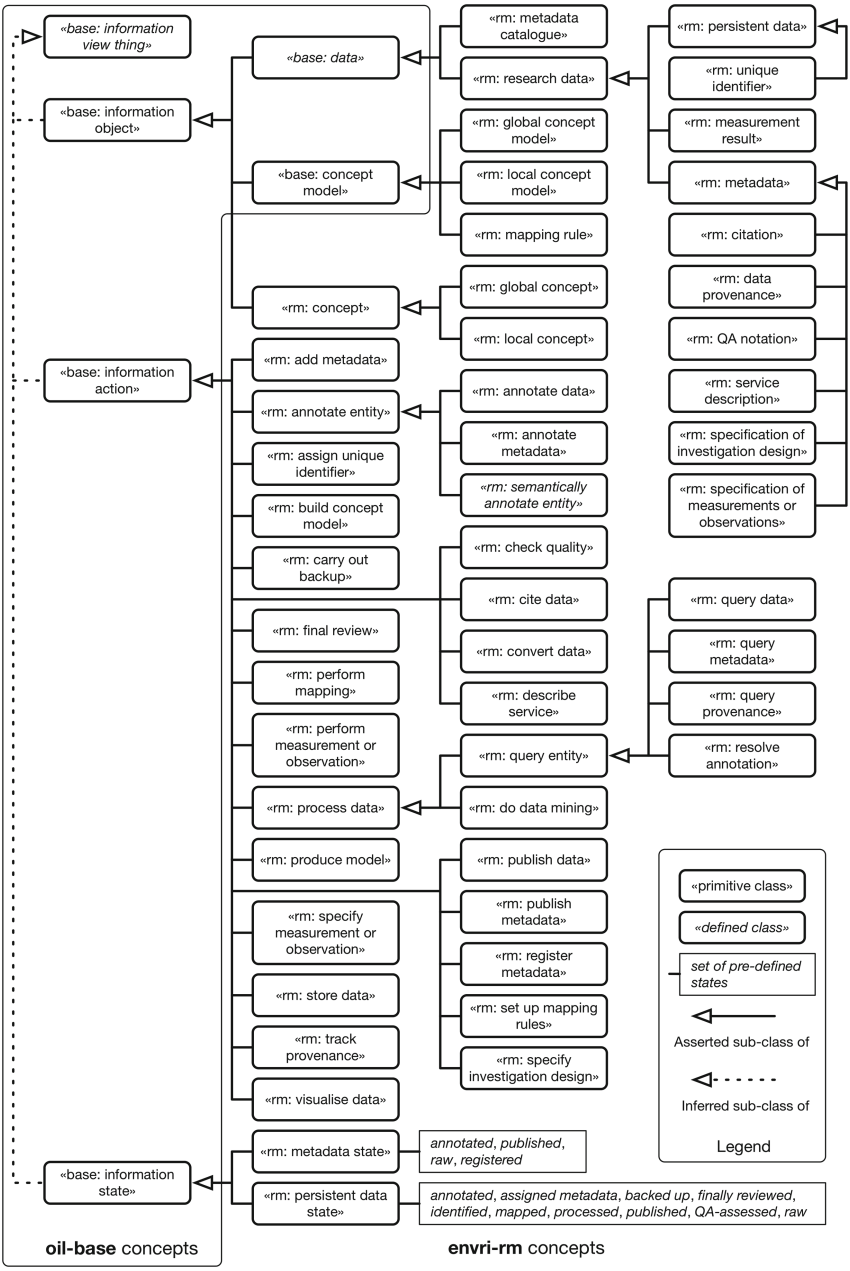


Fig. 3. The concept hierarchy of information viewpoint concepts defined by **envri-rm**. Some defined classes (e.g. for persistent data objects with specific states) have been omitted for brevity.

Defined from the ENVRI RM specification, the **envri-rm** ontology is the primary extension of **oil-base**. This ontology takes the sets of archetypes in each view defined by ENVRI RM, being the classes of object and process considered common across environmental science RIs, and extends OIL-E with concept classes (over 250 at the time of writing) for all of them, allowing better contextualised classification of RI entities and the ability to infer necessary relationships between them. As an example, Fig. 3 shows the set of concepts defined by ENVRI RM for the *information viewpoint*; the concepts are mainly split across sub-classes of information object (e.g. persistent dataset) and information action (e.g. annotate data), with some added information states used to differentiate information objects (e.g. annotated, assigned metadata, backed up or published).

ENVRI RM is an on-going development; with each release of the model, the **envri-rm** ontology must be updated accordingly. Currently, this is done via consultation within the relevant working group in the ENVRI community, based on demand for new stereotypes for RI entities or activities, or discussion regarding the correctness of specific properties or other relationships.

4 Using OIL-E to Model RIs and Research Activities

OIL-E is intended to assist with semantic harmonisation between different RIs by providing a connective ontology for describing RI components and activities based on the archetypes defined by ENVRI RM—essentially to help provide a landscape overview of how RI services are designed and implemented. In particular, OIL-E was designed to be a framework by which we can study how different metadata schemes and controlled vocabularies are used in practice to describe various entities of interest to RIs. Such a study, in the correct context, can be used to expedite alignment and transformation of formal specifications in the service of greater RI interoperability. This entails:

- Comparing different concept models for modelling research assets and data, and identifying commonalities and gaps.
- Building generic tools using existing technologies to handle the search and mapping of models related to RI architecture and specification.

The linking component of OIL-E glues concepts both inside ENVRI RM and between ENVRI RM and external vocabularies. In the latter case, external models can be classified in terms of ENVRI RM in order to help map the landscape of RI-related standards and models. The **envri-rm** ontology only contains a limited set of vocabularies derived from common RI functionality and design patterns, so linking **envri-rm** with external models will also enable domain-specific extensions to ENVRI RM itself. The internal correspondences between the different OIL-E views can potentially be used to indirectly draw associations between concept models with quite different foci (e.g. data versus services).

Notably, OIL-E conflates two major classes of information regarding RIs: schematic information, about the general ‘kinds’ of the element found in a given RI; and instance information, about actual services, datasets, technologies currently found in a RI. Take

for example the Integrated Carbon Observation System (ICOS)[5]. Modelling this RI, we can assert that "ICOS Level 1 data" concerns a general class of dataset found in the ICOS Carbon Portal, the properties of which apply to all instances of such datasets, while there may also be individual examples of Level 1 data product in ICOS that we also model. A description of the former is schematic information, while a description of the latter is instance information. In practice, most OIL-E data so far produced is a mix of schematic information and instance data about invariant parts of RIs. For example, the "ICOS Carbon Portal" is a specific component of the ICOS RI rather than a class of component and thus is instance data, as is the metadata standard "ISO 19139" used for metadata produced by many RIs (though there may also be a class of "ISO 19139 compliant metadata records"). Whether schematic or instance information, the combination of this data provides a description for a RI that can be used to classify not only persistent RI entities such as datasets and services, but also transient events, which (for example) allows such extensions of OIL-E to be used to classify or validate provenance traces.

Information specific to individual RIs is created by providing specific instances of RM archetypes implemented by the RI as well as extending **envri-rm** with concepts particular to the RI, for example as shown in Fig. 4. In this case (which for brevity has been simplified from reality to serve as an exemplar), we extend **envri-rm** for the AnaEE[6] RI (for Analysis and Experimentation in Ecosystems) and show a few of the concepts for AnaEE-specific processes involving the AnaEE metadata catalogue across three views (science, information and computational views). We also show a couple of the specific entities that must be instantiated to support these processes—for example, the AnaEE discovery catalogue service that must invoke all updates to the metadata catalogue in the RI.

RI-specific concepts may apply to any of the views defined by ENVRI RM, with OIL-E providing the vocabulary necessary to relate concepts within and between views. The technology viewpoint of OIL-E, in particular, allows for the identification of specific technologies (i.e. software, hardware and standards) to be linked to particular types and instances of RI datasets and services, which can then be mapped out as knowledge graphs to show how technologies are used and how they relate to various RIs and RI systems, for example as shown in Fig. 5. We can also identify the context in which such technologies are used (e.g. for what kind of dataset or to implement what service) and provide information about where such technologies can be acquired.

It is also possible to extend OIL-E for a specific kind of process rather than a specific RI. Creating a taxonomic model for data quality control (QC) processes is an example of an extension to the base OIL-E ontology that elaborates upon a specific part of RI design. OIL-E defines a class of RI behaviour, *'quality checking behaviour'*, which can be used to classify QC behaviours performed by RIs; an extension for QC processes can enhance that concept by providing a greater range of relations between QC behaviours and other entities representing terms of a taxonomy for QC processes.

[5] https://www.icos-ri.eu/.
[6] https://www.anaee.com/.

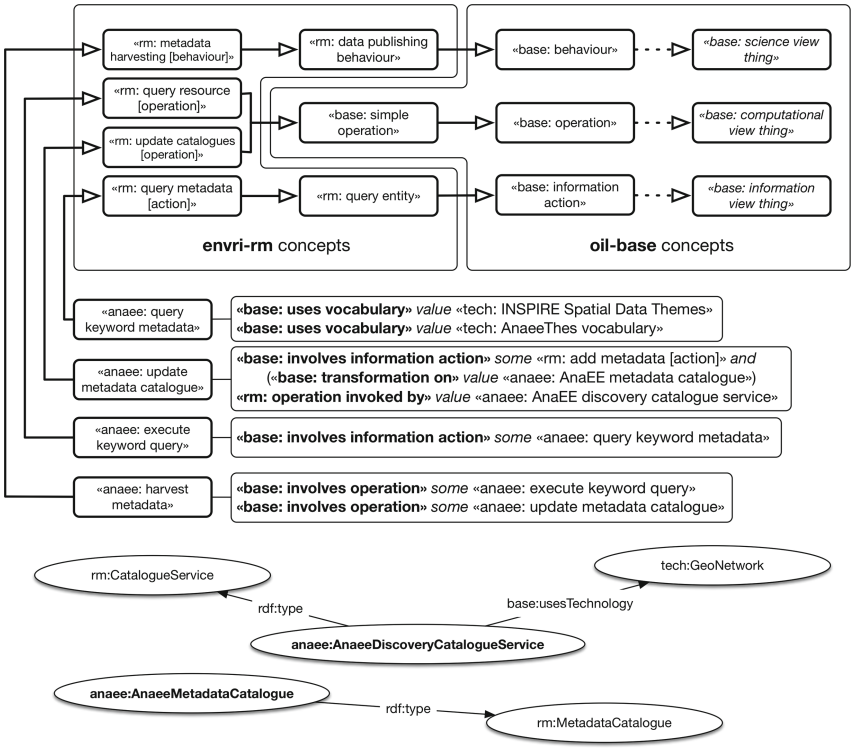


Fig. 4. Extending OIL-E to model components and activities of AnaEE: a (simplified) example of activities involved in the harvesting of metadata for the AnaEE metadata catalogue.

In Fig. 6, a QC process is defined for the EISCAT_3D[7] RI. EISCAT_3D is concerned with using radar observations and the incoherent scatter technique for studies of the atmosphere and near-Earth space above the Arctic. Once fully operational, it will provide a considerable volume of data, in real-time, from its sensor arrays deployed in Norway, Sweden and Finland. These data need to be checked for possible errors or anomalies. As for many research data streams, there need to be multiple phases of quality control to ensure the quality of data reaching researchers. The first of these is described in RDF in Fig. 6. It is performed shortly upon acquisition of new data, is a semi-automated process conducted in real-time by a human technician, performed within the RI itself (since EISCAT_3D acquires the data directly, rather than via intermediaries) and involves a set of activities: statistical checks, corrective measures, technical checks and data enhancements.

There are many other processes defined by ENVRI RM for different parts of the research data lifecycle such as data acquisition or publication. For every such process, OIL-E provides a base stereotype, often with additional requirements for e.g. the actors involved in the process, which can be easily extended with additional controlled vocabulary and sub-concepts using standard Semantic Web technology and techniques.

[7] https://www.eiscat.se/eiscat3d/.

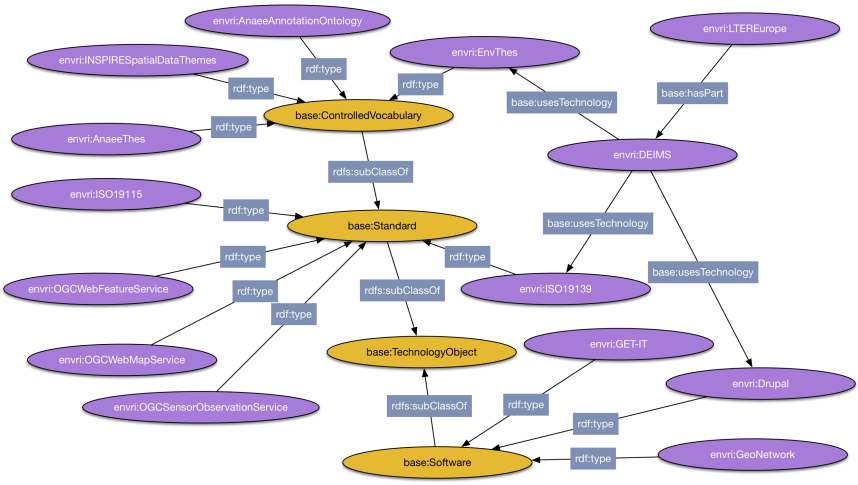


Fig. 5. Linking technologies and standards: the use of different technologies by different RIs can be explored via the knowledge graph generated using RI data in OIL-E.

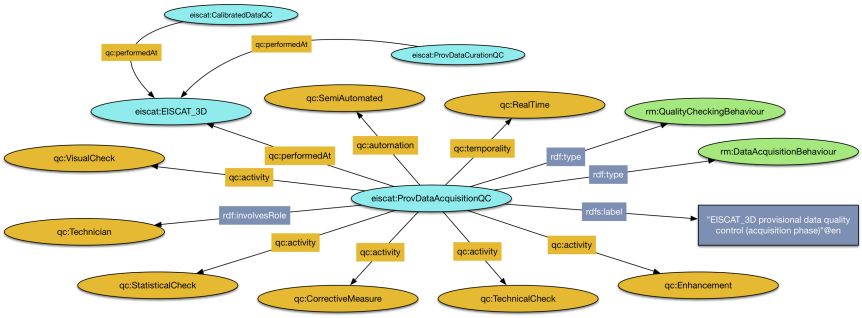


Fig. 6. Modelling quality control processes in OIL-E: example of provisional real-time data quality control on newly-acquired data in EISCAT_3D.

5 The ENVRI Knowledge Base

A key outcome in ENVRIplus that naturally resulted from the creation of OIL-E is the creation of a knowledge base to collect together information about the RIs in the ENVRI community and their activities [12]. The need for such a knowledge base was motivated by the need to better map the semantic landscape of environmental science RIs in Europe, and in particular to gather information about the different metadata schemes, ontologies, thesauri and other controlled vocabularies used by RIs specifically in terms of their application in RI subsystems (as opposed to simply providing another ontology portal).

The ENVRI Knowledge Base in its first iteration as a product of the ENVRIplus project serves three basic purposes:

1. It provides an example of OIL-E in use, providing examples of RI-oriented data structured in accordance with the OIL-E ontologies.
2. It provides a repository for RI architectural information and 'design wisdom' encoded using ENVRI RM that can be programmatically queried and analysed.
3. It serves as a database of information about technologies and standards used by RIs.

The current knowledge base is hosted via a standalone instance of Apache Jena Fuseki[8], which provides a triple store for aggregated RDF data [13] along with a service API and internal reasoning capabilities based on the OWL standard. The knowledge base contains the complete set of OIL-E ontologies along with a representative sample of RI-specific data for the purposes of demonstration and experimentation. Access to the knowledge base is achieved via a SPARQL endpoint [14]. The main landing page for the knowledge base, which also provides a means to try out and modify various sample queries via a Web browser without needing an HTTP/SPARQL request client, can be found via the ENVRI community site at: http://kb.oil-e.net[9].

Figure 7 shows a visualisation of information in the current ENVRI knowledge base as can be viewed by visiting the above landing page. Nodes are colour coded to distinguish concept classes from instance data and data properties, with additional information accessible by directly selecting individual nodes.

When resolving queries sent to it, the knowledge base is able to apply the relations and classifications defined by OIL-E in order to infer results beyond those explicitly asserted in the internal triple store. This allows the full set of ENVRI RM archetypes to be used to guide the discovery and search over all the RI data provided. It should be noted that the scope of the knowledge base as of writing is that of a demonstrator, rather than a production-level system, and so all information about RIs found in the ENVRI knowledge base is provisional, and should not yet be considered an accurate representation of the infrastructures in question. The ENVRI-FAIR project[10] [26], which started January 2019, will build upon this demonstrator to implement a more authoritative knowledge base for the ENVRI community.

For the ENVRI Knowledge Base, we identified four key knowledge capabilities that we wanted to support:

A Survey of the Technical Landscape. The web of knowledge created by semantic linking should help us understand what technologies (including software, standards and vocabularies) are being used by environmental science RIs.

Comparative Solution Analysis. It should be possible to compare solutions developed by environmental science RIs—specifically, given the knowledge of how technologies are used in their proper context, we should be able to compare developments in equivalent contexts.

Gap Analysis and Component Recommendation. Given a reference model for environmental science RIs (i.e. ENVRI RM), it should be possible to identify what is missing in the current development state of a given RI, and based on both that model

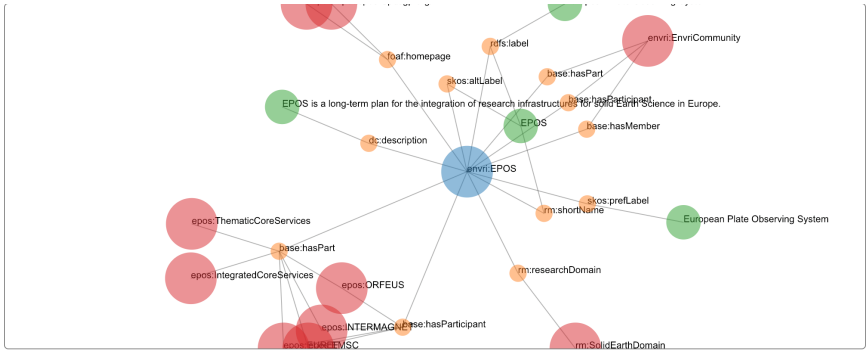
[8] https://jena.apache.org/documentation/fuseki2/.

[9] In the ENVRI-FAIR project, it is planned to be deployed using the ENVRI community domain.

[10] https://www.envri.eu/envri-fair/.

and the solutions developed by other RIs, it should be possible to then make certain recommendations.

Linked Open Research Infrastructure. The web of knowledge created by semantic linking should itself be publicly accessible, machine-navigable, and provide a gateway to the services and data held by the RIs. It should include (where available) data provenance and resource catalogues, and it should (where appropriate) make use of other ENVRI services such as the catalogue service for cross-RI search.



property	data
http://www.w3.org/2000/01/rdf-schema#label	ORFEUS
http://purl.org/dc/elements/1.1/description	ORFEUS is the non-profit foundation to coordinate and promote digital, broadband seismology in the European-Mediterranean area.
http://xmlns.com/foaf/0.1/homepage	http://www.orfeus-eu.org/

Fig. 7. Visualising information in the ENVRI Knowledge Base; showing entities associated with the EPOS research infrastructure.

Given a sufficiently-detailed corpus of information regarding environmental and Earth science RIs backed by a 'standard model' for how such RIs are constructed (i.e. ENVRI RM), it is possible to evaluate individual RIs or RI subsystems in terms of how they compare with similar RIs or against some kind of base criteria. For example, it is possible to compare data quality control processes applied by RIs in the same domain (e.g. marine science) against one another, or against a specific prescribed methodology (e.g. the quality control methodology used for oceanographic data collected by GODAE [15]). Another possibility is to provide tools for RI designers to evaluate their own RIs in terms of compliance with the FAIR data principles [16]. Wilkinson et al. define a number of guidelines for providing Findable, Accessible, Interoperable and Reproducible data that would better support open data science. There are a number of possible approaches that could be taken to evaluate some of these guidelines based on the content of an OIL-E-structured knowledge base:

Findability. Which published data products include globally unique persistent identifiers? Are those identifiers included in the product metadata? What other core metadata does each published product include (or not include)? Does the RI provide an index or

registry for search and discovery of data products? Does it contribute to any external registries?

Accessibility. Can data product metadata be retrieved by a standard, open and free communication protocol, and if so, which one? Does the RI define an authentication and authorisation process for accessing data, and does it use standard, open mechanisms? Are metadata accessible via some means even if the data product described is no longer available?

Interoperability. What data formats, metadata schemes and controlled vocabularies are used to describe/represent (meta)data in the RI? Do those terminological resources comply themselves to the FAIR principles?

Reusability. How rich are the metadata provided for data products? What licences are assigned to the use of data? Is detailed provenance included in the metadata, and does the RI include provenance tracking in its internal processes? Do RI (meta)data meet domain-specific community standards?

Notably, such evaluation does not rely solely on the specification of data products (information view), but also on information about the services provided or delegated by a RI (computational view), the technologies used (technology view) and the general processes defined (science view). Thus, the holistic multi-view specifications permitted by OIL-E using ENVRI RM stereotypes potentially allows for a much more sophisticated analysis of RI status than would be provided by (for example) a catalogue of metadata schemes used by RIs for their primary data products.

6 Discussion

The knowledge base and OIL-E are both the basis for more tools with which to support several useful functions. We can envisage a number of avenues of further development (or in most cases, alignment with existing developments for mutual benefit). These include:

Cross-RI Search and Discovery. OIL-E provides a standard taxonomy for various entities and activities related to RIs, which can be used to classify different kinds of resources as part of a faceted search pipeline. An OIL-E knowledge base can hypothetically act directly as a catalogue service for multiple RIs, but this is not necessarily the best possible approach, as OIL-E is optimised for describing RI design and contextualising RIs' component parts, rather than providing a more traditional metadata scheme for describing RI resources. A knowledge base can be the basis however for a discovery service for heterogeneous research assets (including other catalogues) based on its internal network of relationships based on ENVRI RM, which could conceivably be used to direct queries dispatched to a common search portal to the correct RI resources.

Faster RI Specification Using ENVRI RM. Detailed descriptions of RIs in terms of their architecture, core data products and processes allow for more in-depth investigations and comparisons of RI solutions to various technical problems. ENVRI RM provides the basis for such descriptions, but requires specialist expertise to use effectively, and

has previously been used manually, resulting in the creation of a body of documentation for each RI modelled. OIL-E captures all the key concepts defined by ENVRI RM, and thus a tool based on OIL-E that allows RI architects to more easily specify their RIs using ENVRI RM templates would accelerate the creation of RI data; this data can then be directly inserted into the ENVRI Knowledge Base and used in comparative analyses. The application of standards such as the Shapes Constraint Language (SHACL) [17], to validate data entry into such templates in a way that complements the basic classification capabilities of the OIL-E OWL ontology, would be particularly helpful.

Requirements Recommendation. Using tools such as OIL-E and the ENVRI Knowledge Base, it is possible to do a comparative analysis of the solutions provided by RIs in terms of technology and processes to address various common problems regarding the handling of research data (and other things). This requires a certain degree of constructive analysis of a number of queries. Tools which can interact with the knowledge base on behalf of a user, constructing and interpreting queries behind a more friendly interface, could be very useful for taking full advantage of the corpus of knowledge built up from RI modelling [12].

Provenance Exploration. There are two notable ways in which OIL-E data can interact with provenance data, especially data encoded to the W3C PROV standard [18]. The first is as linking data to various provenance repositories, contextualising the role of the repositories and providing a reference to where the provenance is and how it can be extracted. The second is as a validation framework; given descriptions of RI processes encoded in OIL-E, provenance traces can be checked against those descriptions by mapping agents, entities and activities to the correct OIL-E concepts and then checking whether the relationships described in the provenance trace match those of prescribed by the process model.

Natural Language-Based Document Analysis and Annotation. A significant corpus of existing information about RIs exists in the form of written documentation produced by RI architects and developers. The ability to apply a framework such as OIL-E to annotate uploaded documents, identifying possible references to concepts defined in ENVRI RM in the text, for example, would be useful both to contextualise documents automatically and provide initial descriptions for the RIs and RI components described by the documents. Such descriptions can be verified and extended by human experts, and also used as training data for producing better annotations in future, or perhaps even to identify possible extensions (e.g. new concepts or alternative synonyms for existing concepts) to ENVRI RM. Machine learning tools would thus provide a valuable additional source of data for the knowledge base, or to validate existing models of RIs [27].

The Semantic Web relies on a number of foundational technologies for representing and associating semantics to information, from RDF to OWL and SKOS [19], along with standards for interacting with semantic information (e.g. for search and discovery) such as SPARQL. Considerable attention has been given to the openness, extensibility and computability of such standards, with different options for controlled vocabulary specification depending on the circumstances (e.g. the choice of SKOS over OWL for many vocabulary specification cases [20, 21]). While RI designs could be specified using

something other than Semantic Web technologies (for example based on traditional relational database models), the openness and extensibility of the Semantic Web fit well with the heterogeneity of RI designs and the varying levels of detail in which specific aspects of RI design may or may not be modelled. It should also be noted that RI models are not themselves particularly large in terms of data volume, being constructed of relatively 'high-level' propositions that nonetheless need to be very carefully structured; this also fits the Semantic Web knowledge graph meta-model.

OIL-E's use of RM-ODP (via ENVRI RM) is not wholly new; RM-ODP has been expressed in ontology form as early as 2001 [22]. Applications of ODP have been studied extensively [23], and ODP has been applied to the design of various kinds of infrastructure, including in the Internet of Things (IoT) and Smart Cities contexts [24]. The applicability of ODP, a standard that was developed in the 1990s, to modern concepts of service-oriented architecture and Cloud have been discussed before in research literature [25]. Certainly, the advancement and wide-scale adoption of virtualisation and programmable infrastructure mean that the separation of concerns between the computational and engineering viewpoints (for example) are less clear than they perhaps were originally; modelling a system deployed on a virtual infrastructure and modelling the virtual infrastructure service itself, for instance, would each result in a very different assignment of concepts between the two views. On the other hand, ODP supports the notion of transparencies, the selection of aspects of system design (such as authentication and migration of components) to not be explicitly modelled in specifications so as to reduce confusion, clutter or repetition in design documents. In this light, the explicit acknowledgement that the resources and channels described in the engineering view of a RI specification happen to be virtualised becomes simply another transparency option. Certainly, regardless of whether ODP can be considered to be a sufficiently contemporary specification for the modelling of modern distributed systems, the notion of specifying systems across multiple views is still well-regarded in software engineering research literature.

In many cases, it has been found that most queries on the state of the modelled RI systems focus on a single view, but defining correspondences between views can still be very useful for validating consistency between views of the same RI. Conceivably, different views on the same RI might be maintained by different authorities; the OIL-E multi-view ontology helps keep the different views consistent by identifying expected links between concepts in different views, which RI architects can then evaluate and try to align, either in the description of the RI or, where deficiencies in the subject are identified, in the design of the RI itself. One additional possible extension to OIL-E now being investigated is the integration of SHACL functions into OIL-E. SHACL is a constraint language used to validate RDF graphs and is a refinement of prior de facto standards such as SPIN and SWRL. Unlike OWL it performs closed world validation rather than open-world classification, and also makes the unique name assumption that OWL explicitly does not. SHACL can be used to embed SPARQL queries into RDF graphs as part of rules or functions that can be applied on the content of the graph, providing a means for RI service developers to publish instructions for building (for example) parameterised HTTP requests to their services that other actors can retrieve from the knowledge fabric. Such an approach allows interaction logic to be defined

(and updated) in one place (e.g. the knowledge base or a successor system that may be distributed over several nodes perhaps directly curated by RIs). It also admits the possibility that other information in the linked knowledge graph can be used in a dynamic fashion to introduce some additional interstitial intelligence into the logic.

7 Conclusion

In this chapter, we described how the ENVRI Reference Model (ENVRI RM) was used as the basis for a formal ontology for describing research infrastructure, RI subsystems and RI processes. This ontology, called Open Information Linking for Environmental Research Infrastructures (OIL-E), preserves the multi-view approach of ENVRI RM to provide a flexible framework for RI modelling that can be tailored to the particular interests of different stakeholders in RI design and development; for example to focus on the behaviours of the main actors in a RI, the computational services provided by the RI, or the main datasets curated by the RI. We described the main design principles of OIL-E and how we captured the extensive lexicon of ENVRI RM in a logical concept hierarchy in a way that could be easily applied or extended for specific RIs or particular kinds of activity.

The use of ontologies to capture the vocabulary and relations between entities is a useful means to model information artefacts used in research infrastructure and other knowledge-based systems, but the balance of expressivity and computability poses a continuing challenge. Ontologies are a single tool alongside other forms of schema and rubrics for the capturing of design knowledge and architectural details of the infrastructure. To better explore how best to capture RI design information in a way that is machine-readable, programmatically queryable, and above all *useful*, we applied OIL-E as the underlying structure for a knowledge base of European environmental and Earth science RIs participating in the ENVRI community. This knowledge base was created to demonstrate how an information corpus for RIs might be used to analyse and compare RI designs, as well as to document the technologies, software and standards used by RIs in their appropriate operational contexts. We reviewed the current state of the knowledge base, and discussed a number of ways in which it can be improved to permit the handling of a greater range of queries of possible interest to RI designers.

The next major objective of the ENVRI community is to facilitate the adoption of the FAIR data principles for research data gathered in the atmospheric, marine, solid earth and biodiversity domains, and to develop sustainable FAIR data services for research communities as part of the broader push towards better open data science and more seamless interoperability between different data providers. Further development of the ENVRI Knowledge Base would greatly support this effort by providing a clear means for RI developers to evaluate their RIs' progress towards greater 'FAIRness' and to explore the technology choices made by their fellow developers in other scientific domains. Such development has been committed to as part of the next phase of ENVRI, the ENVRI-FAIR project, which began work in January 2019. The main priorities in knowledge base development will thus be to take the lessons and data prototypes developed in the previous ENVRIplus project in order to create a more extensive, complete, 'production-level' knowledge resource.

Acknowledgements. This work was supported by the European Union's Horizon 2020 research and innovation programme via the ENVRIplus project under grant agreement No 654182.

References

1. Zhao, Z., et al.: Reference model guided system design and implementation for interoperable environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science, pp. 551–556. IEEE, Munich (2015). https://doi.org/10.1109/eScience.2015.41
2. Nieva de la Hidalga, A., et al.: The ENVRI Reference Model (ENVRI RM) version 2.2, November 2017. https://doi.org/10.5281/zenodo.1050349
3. Martin, P., et al.: Open information linking for environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science (e-Science), pp. 513–520. IEEE (2015). http://dx.doi.org/10.1109/eScience.2015.66
4. Martin, P., Chen, Y., Hardisty, A., Jeffery, K., Zhao, Z.: Computational challenges in global environmental research infrastructures (Chap. 12). In: Chabbi, A., Loescher, H.W. (eds.) Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities, pp. 305–340. CRC Press, Boca Raton (2017). https://zenodo.org/record/3361569
5. ISO 10746-1: Information technology—Open Distributed Processing—Reference Model: Overview. ISO/IEC standard. International Organization for Standardization (1998)
6. Linington, P.F., Milosevic, Z., Tanaka, A., Vallecillo, A.: Building Enterprise Systems with ODP: An Introduction to Open Distributed Processing. CRC Press, Boca Raton (2011)
7. Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., Lewis, S.E.: The environment ontology: contextualising biological and biomedical entities. J. Biomed. Semant. **4**(1), 43 (2013)
8. Arp, R., Smith, B., Spear, A.D.: Building Ontologies with Basic Formal Ontology. The MIT Press, Cambridge (2015)
9. Jörg, B.: CERIF: the common European research information format model. Data Sci. J. **9**, 24–31 (2010)
10. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
11. W3C OWL Working Group: OWL 2 web ontology language. W3C recommendation, W3C (2012). https://www.w3.org/TR/2012/REC-owl2-overview-20121211/
12. Zhao, Z., et al.: Knowledge-as-a-service: a community knowledge base for research infrastructures in environmental and earth sciences. In: 2019 IEEE World Congress on Services (SERVICES), pp. 127–132. IEEE, Milan (2019). https://doi.org/10.1109/SERVICES.2019.00041
13. Wood, D., Cyganiak, R., Lanthaler, M.: RDF 1.1 concepts and abstract syntax. W3C recommendation, W3C (2014), http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/
14. W3C SPARQL Working Group: SPARQL overview. W3C recommendation, W3C. http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/
15. Cummings, J.A.: Ocean data quality control. In: Schiller, A., Brassington, G. (eds.) Operational Oceanography in the 21st Century, pp. 91–121. Springer, Dordrecht (2011)
16. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**, 1–9 (2016). https://doi.org/10.1038/sdata.2016.18
17. Kontokostas, D., Knublauch, H.: Shapes constraint language (SHACL). W3C recommendation, W3C, July 2017. https://www.w3.org/TR/2017/REC-shacl-20170720/
18. Groth, P., Moreau, L.: PROV-overview. W3C note, W3C (2013). http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

19. Bechhofer, S., Miles, A.: SKOS simple knowledge organization system reference. W3C recommendation, W3C (2009). http://www.w3.org/TR/2009/REC-skos-reference-20090818/
20. Stellato, A.: Dictionary, thesaurus or ontology? Disentangling our choices in the semantic web jungle. J. Integr. Agric. **11**(5), 710–719 (2012)
21. Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., Summers, E.: Key choices in the design of simple knowledge organization system (SKOS). Web Semant.: Sci. Serv. Agents World Wide Web **20**, 35–49 (2013)
22. Wegmann, A., Naumenko, A.: Conceptual modelling of complex systems using an RM-ODP based ontology. In: 2001 Proceedings of the Fifth IEEE International Enterprise Distributed Object Computing Conference. EDOC 2001, pp. 200–211. IEEE (2001)
23. Kilov, H., Linington, P.F., Romero, J.R., Tanaka, A., Vallecillo, A.: The reference model of open distributed processing: foundations, experience and applications. Comput. Stand. Interfaces **35**(3), 247–256 (2013)
24. Román, I., Madinabeitia, G., Jimenez, L., Molina, G., Ternero, J.: Experiences applying RM-ODP principles and techniques to intelligent transportation system architectures. Comput. Stand. Interfaces **35**(3), 338–347 (2013)
25. Jebbar, M., Sekkaki, A., Benamar, O.: Integration of SOA and cloud computing in RM-ODP. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 97–105. IEEE (2012)
26. Petzold, A., et al.: ENVRI-FAIR - interoperable environmental FAIR data and services for society, innovation and research. In: 15th IEEE International Conference on e-Science, San Diego, US (2019). https://doi.org/10.1109/escience.2019.00038, https://zenodo.org/record/3462816
27. Liao, X., Zhao, Z.: Unsupervised approaches for textual semantic annotation, a survey. ACM Comput. Surv. **52**(4), 45 (2019). https://doi.org/10.1145/3324473. Article 66

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.